ASYMLORA: UNLOCKING THE POWER OF MULTIMODAL LLMS VIA ASYMMETRIC LORA

Xuyang Wei[1], **Chunlin Tian**[2], **Li Li**[2]
[1]University of Electronic Science and Technology of China
[2]University of Macau
7yun1uxdjj@gmail.com, tianclin0212@gmail.com, LLiLi@um.edu.mo

ABSTRACT

Effective instruction fine-tuning on diverse image-text datasets is crucial for developing a versatile Multimodal Large Language Model (MLLM), where dataset composition dictates the model's adaptability across multimodal tasks. However, complex datasets often contain inherent conflicts—stemming from modality-specific optimization objectives—and latent commonalities that enable cross-task transfer, which most existing approaches handle separately. To bridge this gap, we introduce AsymLoRA, a parameter-efficient tuning framework that unifies knowledge modularization and cross-modal coordination via asymmetric LoRA: task-specific low-rank projections (matrix B) that preserve distinct adaptation pathways for conflicting objectives, and a shared projection (matrix A) that consolidates cross-modal commonalities. Extensive evaluations demonstrate that AsymLoRA consistently surpasses both vanilla LoRA, which captures only commonalities, and LoRA-MoE, which focuses solely on conflicts, achieving superior model performance and system efficiency across diverse benchmarks. GitHub.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Huang et al., 2023; Liu et al., 2023; Zhu et al., 2023) integrate pre-trained vision encoders with LLMs, enabling models to comprehend and generate responses based on both visual and textual inputs. To enhance their ability to handle diverse modalities and downstream tasks, MLLMs leverage instruction tuning, where models are fine-tuned on multimodal instruction-following dialogues synthesized from diverse multimodal tasks. Parameter-Efficient Fine-Tuning (PEFT) techniques (Houlsby et al., 2019; Liu et al., 2021), such as Low-Rank Adaptation (LoRA) (Hu et al., 2021), enhance adaptability by injecting small trainable components into the model, significantly reducing trainable parameters while preserving or even improving task performance. However, directly applying LoRA in multi-task learning (see Figure 1 (a)) can lead to conflicting optimization objectives, where task-specific adaptations interfere with each other, potentially canceling out useful task-specific updates. To address this, Mixture-of-Experts (MoE) (Jacobs et al., 1991; Jordan & Jacobs, 1994) extends LoRA by introducing specialized modules (see Figure 1 (b)) that learn task-specific knowledge (Lin et al., 2024; Chen et al., 2024), improving alignment across diverse modalities. However, multi-task datasets inherently contain both conflicts—arising from modality-specific optimization goals—and latent commonalities that facilitate cross-task transfer (Tian et al., 2024; Feng et al., 2023), which cannot be effectively tackled in isolation.

In this work, We propose AsymLoRA, an asymmetric LoRA architecture designed for MLLM instruction fine-tuning. Unlike vanilla LoRA, which applies a single pair of low-rank decomposed matrices to the Transformer MLP layers, AsymLoRA introduces a shared A matrix for capturing common knowledge and task-specific B matrices for independent task adaptation (see Figure 1 (c)). During inference, AsymLoRA dynamically selects task-specific B matrices in a MoE manner, effectively balancing commonalities and conflicts across multi-task datasets to enhance both efficiency and performance. Through extensive experiments on diverse data configurations, we demonstrate that AsymLoRA effectively mitigates conflicts between instruction datasets while leveraging their shared knowledge, achieving superior performance and efficiency with fewer parameters. We summarize our primary contributions as follows:

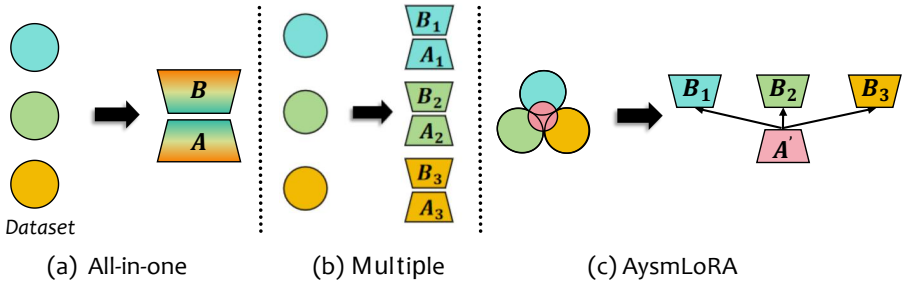

Figure 1: Illustration of LoRA architecture changes in AsymLoRA. (a) Vanilla LoRA applies a single shared adaptation across all tasks, relying solely on common synergies but introducing conflicts during fine-tuning. (b) Multiple task-specific LoRA modules mitigate interference by isolating tasks but focus only on differences, limiting generalization and increasing overhead. (c) AsymLoRA introduces an asymmetric structure with a shared A matrix for common knowledge and multiple B matrices for task-specific features, balancing generalization and efficiency.

- Based on an advanced MLLM model and large scale datasets, dentify the inherent conflicts and commonalities in instruction fine-tuning MLLMs on mixtures of multimodal instruction datasets.
- We propose AsymLoRA, an asymmetric LoRA architecture that addresses conflicts through task-specific B matrices while capturing commonalities via a shared A matrix.
- Extensive experiments across multiple benchmarks validate that AsymLoRA consistently outperforms both vanilla LoRA and LoRA-MoE in terms of both performance and efficiency across various data configurations.

2 Background and Motivation

Low-Rank Adaptation (Hu et al., 2021) is an efficient fine-tuning technique for large pre-trained models, introducing small low-rank matrices (A and B) that can be applied to arbitrary linear layers. Formally, for a linear transformation $h = Wx$ with input $x \in \mathbb{R}^{d_i}$ and weight $W \in \mathbb{R}^{d_o \times d_i}$, LoRA learns a low-rank decomposed update:

$$y' = y + \Delta y = Wx + BAx \tag{1}$$

where $y \in \mathbb{R}^{d_o}$ is the output, and $A \in \mathbb{R}^{r \times d_i}$, $B \in \mathbb{R}^{d_o \times r}$ are low-rank matrices with $r \ll \min(d_o, d_i)$ as the chosen rank. Typically, B is initialized to zeros, while A follows Kaiming Uniform initialization (He et al., 2015). During fine-tuning, only A and B are updated, keeping the original model parameters frozen, thus significantly reducing computational overhead.

Observation I: Knowledge contains inherent conflicts. Instruction fine-tuning on diverse image-text datasets is crucial for enhancing the performance of MLLMs, where the configuration of training data plays a pivotal role. However, we observe that when instruction data from different domains are combined, inherent conflicts between domain-specific optimization objectives become inevitable. These conflicts often lead to a significant performance drop in certain domains compared to fine-tuning on a single-domain dataset. As shown in Table 1, we fine-tune LLaVA (Liu et al., 2023) using instruction data from two distinct domains—Visual Question Answering (VQA) (Antol et al., 2015) and Generative (Gen, including LLaVA-15k (Liu et al., 2023), VQG (Mostafazadeh et al., 2016))—both separately and in combination. While in certain cases, such as the "Yes/No" (78.08%) and "Other" (36.41%) tasks in VizWiz, the increased data volume from mixing domains leads to performance gains, benchmark evaluations generally reveal that naively combining instruction data from different domains significantly degrades performance. For instance, on the TextQA Singh et al. (2019) test set, the Gen-only model achieves 54.25%, whereas the VQA+Gen mixed model drops to 43.25%. These results underscore that the scalability of multi-domain instruction tuning is inherently constrained by dataset conflicts, highlighting the need for more sophisticated strategies to ensure effective multi-domain adaptation.

Schemes	TextVQA (%)	VizWiz(%)				MME	
		Other	Unanswerable	Yes/No	Number	Perception	Cognition
Single LoRA							
VQA	38.08	31.81	39.19	73.64	22.76	1152.46	224.64
Gen	54.80	27.13	76.32	73.9	30.24	1255.63	296.07
VQA+Gen	43.25	36.41	28.91	78.08	22.28	1203.66	268.92
Multiple LoRA (MoE)							
VQA+Gen	54.05	39.66	24.8	82.15	33.98	1454.37	324.64

Table 1: Performance of fine-tuning MLLM (LlaVA-1.5-7B (Liu et al., 2023)) on benchmarks (TextVQA(Singh et al., 2019), VizWiz(Bigham et al., 2010), MME(Fu et al., 2024)) across different instruction data domains.

Observation II: Knowledge contains latent commonalities While modularity is essential, knowledge across different tasks is often complementary, allowing shared learning to enhance performance. MLLMs should be capable of capturing, integrating, and evolving with diverse knowledge from multiple sources and perspectives, enabling collaborative learning across various domains. As shown in Table 1, we fine-tune separate LoRA modules on different instruction datasets and treat each LoRA module as an expert, dynamically combining them via Mixture-of-Experts (MoE) based on randomized inputs. MoE-LoRA achieves the highest accuracy in certain cases, such as MME scores for Perception (1454.37) and Cognition (324.64), as well as the "Other" (39.66%) and "Yes/No" (82.15%) categories in the VizWiz benchmark. These results demonstrate that LoRA-MoE effectively leverages modularity by using task-specific modules, reducing interference, and enhancing performance. However, in other scenarios, MoE-LoRA underperforms compared to single-domain fine-tuning. For example, on TextVQA, it falls short of the Gen-only fine-tuned model, and in the Unanswerable task of VizWiz, it even records the worst performance. This suggests that focusing solely on task differences while overlooking latent commonalities can limit the full potential of the data, ultimately reducing model effectiveness. A balanced approach that respects both task-specific adaptations and shared knowledge is necessary for optimal performance.

3 METHODOLOGY

3.1 ASYMLORA ARCHITECTURE

As illustrated in Figure 2, AsymLoRA introduces an asymmetric design for efficient MLLM instruction fine-tuning, addressing the limitations of traditional symmetric LoRA approaches. Unlike conventional methods that apply both A and B matrices uniformly across all tasks, AsymLoRA maintains a shared low-rank matrix A to capture common knowledge while introducing task-specific low-rank matrices B_i to enable specialized adaptation. Formally, given a dataset $D = \{D_1, D_2, \ldots, D_N\}$ where each D_i corresponds to a subtask T_i, our objective is to optimize shared parameters A and task-specific parameters B_i to minimize the task-specific loss L_i for each T_i:

$$\min_{A,B_i} \sum_{i=1}^{N} L_i(T_i; A, B_i). \tag{2}$$

The shared matrix A facilitates knowledge transfer across tasks, reducing the number of trainable parameters and enhancing generalization, while the task-specific B_i matrices provide targeted adaptations, mitigating interference between conflicting task objectives.

3.2 MIXTURE OF ASYMLORA EXPERTS

To further enhance adaptability and performance in multi-task learning, we extend AsymLoRA with a Mixture of Experts (MoE) mechanism. In this approach, multiple experts share a common low-rank matrix A, representing global knowledge across tasks, while each expert is associated with a distinct set of task-specific matrices $\{B_1^j, B_2^j, \ldots, B_N^j\}$, where j denotes the expert index. Given a task T_i, a gating network dynamically selects the most suitable expert based on task-specific input features. The gating network assigns a weight w_j to each expert j and computes the final task-

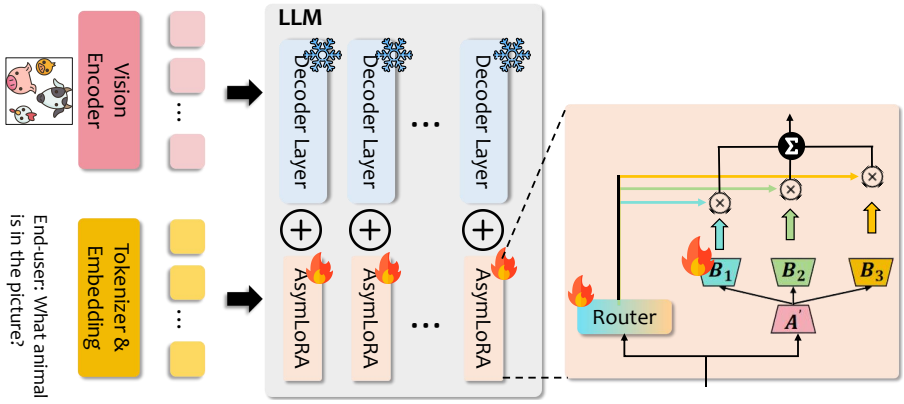

Figure 2: Architecture and workflow of AsymLoRA. The shared low-rank matrix A captures global knowledge across tasks, while task-specific low-rank matrices B_i enable independent adaptation for each task

specific transformation as:

$$\text{Output}_i = f\left(x; A, \sum_j w_j B_i^j\right), \tag{3}$$

This design ensures that while the shared matrix A provides consistent general knowledge, the expert-specific B_i^j matrices enable flexible task-specific adaptation. By dynamically selecting the most appropriate expert, AsymLoRA with MoE improves the model's ability to handle diverse tasks while minimizing interference, leading to more robust and efficient multi-task learning.

4 Experiments

4.1 Experiment Setting

Model Following LLaVA-1.5 (Liu et al., 2023) utilizing CLIP ViT-L (Radford et al., 2021) as the vision encoder with a 336×336 input resolution and a 14×14 patch size. A two-layer MLP adapter processes the 576 tokens extracted from ViT. The language model is Vicuna-7B (Chiang et al., 2023), with both ViT and Vicuna weights kept frozen throughout training. Unless stated otherwise, LoRA is applied to the LLM with a rank of 32, the number of B matrix is initialized N = 3.

Dataset and Benchmarks We evaluate our model in single-domain and multi-domain settings across diverse multimodal tasks. 1) For the single-domain setting, training is conducted on **Conversation_58k**, a dataset with 58,000 conversational examples for dialogue-based learning, and **LLaVA_v1_5_mix665k**, a large-scale mixed dataset for multimodal training. 2) In the multi-domain setting, training combines **VQA, LLaVA-15k, and VQG**, where VQA is a large-scale dataset for open-ended visual question answering, LLaVA-15k consists of 15,000 vision-language task samples, and VQG facilitates natural question generation for conversational AI. Evaluation is performed on multiple benchmarks, including **MME** (Fu et al., 2024) (multimodal integration and reasoning), **GQA** (Hudson & Manning, 2019) (scene graph-based VQA), **MM-Vet** (Yu et al., 2024) (Integrated Capabilities of MLLM) **VizWiz** (Bigham et al., 2010) (real-world VQA with noisy images), and **TextVQA** (Singh et al., 2019) (requiring integration of textual and visual information).

4.2 Overall Performance

We present the experimental results of AsymLoRA and competing baselines across three evaluation settings: single-domain conversation tasks (Table 2), single-domain general tasks (Table 3), and multi-task domain settings (Table 4). The results demonstrate that AsymLoRA consistently outperforms all other schemes, validating its effectiveness in multimodal instruction fine-tuning.

Performance in Single-Domain Tasks In Table 2, AsymLoRA achieves a TextVQA score of 55.51%, surpassing MoE-LoRA (53.33%) and significantly outperforming LoRA (36.43%), indicating its superior ability to integrate textual and visual cues. On the MME benchmark, AsymLoRA achieves the highest Perception (1327.93) and Cognition (287.14) scores, outperforming MoE-LoRA (1121.88, 270.01) and LoRA (911.3, 278.21), showcasing its enhanced multimodal reasoning and feature extraction capabilities. Furthermore, in GQA, AsymLoRA attains the highest accuracy (59.60%) while minimizing distribution shift (1.50), highlighting its robust generalization in structured reasoning tasks. Similarly, in Table 3, AsymLoRA achieves consistent gains in the VizWiz benchmark, leading in Unanswerable (81.70%) and Number (43.33%), with the highest overall average score (51.31%), demonstrating its robustness in real-world visual question answering. This improvement is attributed to AsymLoRA's asymmetric design, which effectively balances common knowledge (A matrix) and task-specific adaptation (B matrices), mitigating conflicts between different domains while retaining the strengths of both LoRA and MoE-LoRA.

Robust Multi-Task Adaptation In the multi-task domain setting (Table 4), AsymLoRA continues to outperform competing methods across Perception, Cognition, and reasoning-based benchmarks, demonstrating its superior capability in handling diverse multimodal challenges. Specifically, it achieves the highest TextVQA score (54.25%) and VizWiz average (38.10%), improving upon MoE-LoRA (53.84%, 37.44%) and LoRA (43.25%, 36.00%). These results highlight AsymLoRA's advantage in adapting dynamically to different domains while preserving effective knowledge transfer across tasks, leading to overall superior multi-task performance.

AsymLoRA consistently outperforms LoRA and MoE-LoRA by effectively resolving conflicts in multi-task instruction tuning while leveraging task-specific adaptations. Its asymmetric architecture enables more efficient knowledge sharing and specialization, making it a highly effective fine-tuning approach for multimodal learning.

Schemes	**TextVQA** (%)	**MM-Vet(%)**	**MME**		**GQA**			
			Perception	**Cognition**	**Binary** (%)	**Open** (%)	**Acc.** (%)	**Dis.** (↓)
LoRA	36.43	<u>31.5</u>	911.3	<u>278.21</u>	12.04	8.32	10.03	3.62
MoE-LoRA	<u>53.33</u>	31.2	<u>1121.88</u>	270.01	<u>66.90</u>	<u>42.73</u>	<u>53.82</u>	<u>1.58</u>
AsymLoRA	**55.51**	**31.8**	**1327.93**	**287.14**	**75.23**	**46.35**	**59.60**	**1.50**

Table 2: Evaluation results for fine-tuning LlaVA on single domain (Conversation-58k).

Schemes	**VizWiz (%)**					**MME**	
	Unanswerable	**Yes/No**	**Number**	**Other**	**Average**	**Perception**	**Cognition**
LoRA	55.32	75.27	<u>39.43</u>	**39.22**	45.12	<u>1446.44</u>	262.5
MoE-LoRA	<u>72.47</u>	**75.79**	39.27	<u>38.75</u>	<u>49.43</u>	1446.31	<u>306.07</u>
AsymLoRA	**81.70**	<u>75.53</u>	**43.33**	37.79	**51.31**	**1489.19**	**367.14**

Table 3: Evaluation results for fine-tuning LlaVA on single domain (LLaVA-mix665k).

Schemes	**VizWiz (%)**					**TextVQA (%)**
	Unanswerable	**Yes/No**	**Number**	**Other**	**Average**	
LoRA	**28.91**	78.08	22.28	36.41	36.00	43.25
MoE-LoRA	<u>24.80</u>	**82.15**	<u>33.98</u>	<u>39.66</u>	<u>37.44</u>	<u>53.84</u>
AsymLoRA	23.65	<u>78.10</u>	**34.80**	**41.35**	**38.10**	**54.25**

Table 4: Evaluation results for fine-tuning LlaVA on multi-task domain (VQA, LLaVA-15k, VQG).

5 CONCLUSION

This work presents AsymLoRA, a parameter-efficient tuning framework that balances modality-specific conflicts and cross-task commonalities in MLLM fine-tuning. By leveraging task-specific B matrices for adaptation and a shared A matrix for knowledge transfer, AsymLoRA outperforms vanilla LoRA and LoRA-MoE, achieving superior performance and efficiency across benchmarks. These results highlight its effectiveness as a scalable solution for multi-task instruction fine-tuning.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 333–342, 2010.

Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. *arXiv preprint arXiv:2305.09955*, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL `https://arxiv.org/abs/2306.13394`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 2021.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Cheng-Zhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *arXiv preprint arXiv:2401.16160*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.